

**HHS PUBLIC ACCESS**

Author manuscript

Biometrika. Author manuscript; available in PMC 2015 November 02.

Published in final edited form as:

Biometrika. 2014 June ; 101(2): 477–483. doi:10.1093/biomet/asu006.

Hypothesis testing for band size detection of high-dimensional banded precision matrices

BAIGUO AN,School of Statistics, Capital University of Economics and Business, Beijing 100070, China,
anbg200@gmail.com**JIANHUA GUO,** andSchool of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin 130024,
China, jhguo@nenu.edu.cn**YUFENG LIU**Department of Statistics and Operations Research, University of North Carolina, Chapel Hill,
North Carolina 27599, U.S.A., yfliu@email.unc.edu

Abstract

Many statistical analysis procedures require a good estimator for a high-dimensional covariance matrix or its inverse, the precision matrix. When the precision matrix is banded, the Cholesky-based method often yields a good estimator of the precision matrix. One important aspect of this method is determination of the band size of the precision matrix. In practice, crossvalidation is commonly used; however, we show that crossvalidation not only is computationally intensive but can be very unstable. In this paper, we propose a new hypothesis testing procedure to determine the band size in high dimensions. Our proposed test statistic is shown to be asymptotically normal under the null hypothesis, and its theoretical power is studied. Numerical examples demonstrate the effectiveness of our testing procedure.

Keywords

Band size; Banded matrix; Cholesky decomposition; Covariance matrix; High-dimensional hypothesis test; Multiple comparison; Precision matrix

1. Introduction

Techniques such as thresholding and banding have been widely used to achieve accurate covariance matrix estimation. The thresholding method sets small elements of the sample covariance matrix to zero to obtain more stable estimators (Bickel & Levina, 2008a; Rothman et al., 2009). Another approach is to perform banding. In many applications, random variables have a natural ordering, in which case it is often assumed that two random variables that are far apart may have weak correlation. In this setting, one approach is to

Supplementary material

Supplementary material available at *Biometrika* online includes additional simulation results, numerical comparisons, further details on the real-data examples, and proofs of the theorems.

band or taper the sample covariance matrix (Bickel & Levina, 2008b; Cai et al., 2010). Since these estimators may not be positive definite, Rothman et al. (2010) proposed a Cholesky-based covariance regularization method to ensure positive definiteness.

In practice, the inverse covariance matrix, also known as the precision matrix, may be of primary interest. When the data are multivariate Gaussian, the precision matrix can be used to infer the conditional dependence structure of random variables. One popular method for precision matrix estimation is to maximize the Gaussian likelihood with a penalty on the precision matrix (Meinshausen & Buhlmann, 2006; Yuan & Lin, 2007; Friedman et al., 2008).

When the variables of interest have a natural order, it is often assumed that two random variables are not partially correlated when the distance between them is large enough. This is equivalent to assuming that the precision matrix is banded with a band size K_0 . In such situations, the relationships between the random variables can be represented using a series of simple linear regression equations through the Cholesky decomposition for the covariance matrix. The number of parameters in each regression equation is no greater than $K_0 + 1$. Usually K_0 is smaller than the sample size, so the regression parameters can be estimated well in high-dimensional small-sample-size cases. Once the estimators for those regression parameters are obtained, one can find an estimator for the precision matrix through the modified Cholesky decomposition (Pourahmadi, 1999; Wu & Pourahmadi, 2003; Huang et al., 2006; Bickel & Levina, 2008b). The band size K_0 is unknown in practice and is often determined by crossvalidation, but this can be time-consuming. Furthermore, crossvalidation can be quite unstable, and may not identify the true band size K_0 when K_0 is relatively large. Consequently, it is necessary to have an alternative method for determining the band size K_0 more accurately.

In this paper, we propose a hypothesis testing procedure to determine the band size of high-dimensional precision matrices. Our null hypothesis states that the band size is smaller than a prespecified constant, and the alternative is that the band size is greater than or equal to that constant. We show that under this null hypothesis our proposed test statistic is asymptotically normal, and we study the power of the test using our proposed statistic. Our numerical studies confirm the theoretical findings and show that our testing procedure can correctly identify the band size of a banded precision matrix in high-dimensional cases.

2. Methodology

2.1. Notation and the modified Cholesky decomposition

Let X_i ($i = 1, \dots, n$) be the observation collected from the i th subject. Here, the

$X_i = (X_{i1}, \dots, X_{ip})^T \in \mathbb{R}^p$ ($i = 1, \dots, n$) are p -dimensional vectors which are independent and normally distributed with mean zero and covariance matrix Σ . The modified Cholesky decomposition of Σ is denoted by $\Sigma = LDL^T$, where L is a lower triangular matrix with ones on the diagonal and $D = \text{diag}(d_1, \dots, d_p)$ is a diagonal matrix. Write L^{-1} as T , which is also lower triangular with ones on the diagonal. Then the precision matrix $\Omega = \Sigma^{-1} = (\omega_{ij})$ can be written as $\Omega = T^T D^{-1} T$.

To estimate Σ , it suffices to estimate $T = (t_{ij})$ and D . Let $\varepsilon_i = T X_i$. Then $X_{i1} = \varepsilon_{i1}$ and, for every $j > 1$, $X_{ij} = \sum_{q=1}^{j-1} (-t_{jq}) X_{iq} + \varepsilon_{ij}$. The elements of ε_i are independent and normally distributed with mean zero, and the covariance matrix of the ε_i is D . Given the data, one can fit the above regression equations to estimate T and D . However, with large p and small n , estimators for T and D obtained through fitting these regression equations may not work well, and so it is common to impose some kind of regularization on T and D (Huang et al., 2006; Levina et al., 2008).

When the variables of interest have a natural order and the banded assumption on Ω holds, Rothman et al. (2010) showed that the band size of Ω is K_0 if and only if the corresponding Cholesky factor T is also banded with band size K_0 . In fact, the model considered here is a special case of chordal graphical models. As was shown in a 2005 unpublished technical report by J. Dahl et al. from the University of California, Los Angeles, for a general chordal graph, the Cholesky factor T still has the same sparse structure as Ω after reordering the variables using a perfect elimination order. Hence, under this banded assumption on Ω , the regression equations can be rewritten as $X_{i1} = \varepsilon_{i1}$, and for every $j > 1$ with $(j - K_0)_1 = \max\{1, j - K_0\}$,

$$X_{ij} = \sum_{q=(j-K_0)_1}^{j-1} (-t_{jq}) X_{iq} + \varepsilon_{ij}. \quad (1)$$

If K_0 is small, then fitting the above regression equations can give a good estimator for T and D in this large p , small n setting. Given estimators of T and D , the estimator for Ω can be derived accordingly. Therefore, correctly identifying K_0 is very important in practice. Wu & Pourahmadi (2003) suggested using the AIC or BIC criteria to choose K_0 , but these may not work well in high dimensions. Crossvalidation is often used to estimate K_0 (Bickel & Levina, 2008b), but it is computationally intensive and can be very unstable.

2.2. Test statistic

To determine the true band size K_0 , we consider the hypothesis testing problem

$$H_0: K \leq k - 1 \quad \text{versus} \quad H_1: K \geq k, \quad (2)$$

where K denotes the band size parameter, whose true value is K_0 , and k is a prespecified

positive number smaller than $n - 4$. Let $t_j^{(k)} = (t_{j,(j-k)_1}, \dots, t_{j,j-1})^T$. Define

$\mathcal{X} = (X_1, \dots, X_n)^T$. Denote the j th column of \mathcal{X} by \mathcal{X}_j and define

$\mathcal{X}_j^{(k)} = (\mathcal{X}_{(j-k)_1}, \dots, \mathcal{X}_{j-1})$. By fitting the regression equation (1), one can obtain the

estimator for $t_j^{(k)}$ as $\hat{t}_j^{(k)} = - \left(\mathcal{X}_j^{(k)T} \right)^{-1} \mathcal{X}_j^{(k)T} \mathcal{X}_j$ and the estimator for d_j as

$\hat{d}_j^{(k)} = \|\mathcal{X}_j + \mathcal{X}_j^{(k)} \hat{t}_j^{(k)}\|^2 / [n - \{j - (j - k)_1\}]$. When H_0 is true, we know that $t_{j,j-k} = 0$ for every $j > k$, and the corresponding estimators should be small. Hence we construct the test

statistic $L = \sum_{j=k+1}^p \hat{t}_{j,j-k}^{(k)2}$.

From the above analysis, one can reject H_0 when L is sufficiently large. To have an accurate testing procedure, the exact or asymptotic distribution of L under the null hypothesis H_0 is needed, and the Type I error must be controlled accordingly. Moreover, the variances of the $\hat{t}_{j,j-k}^{(k)2}$ for $k+1 \leq j \leq p$ may be different from one another, and hence the derivation of the asymptotic distribution of L under the null hypothesis can be challenging. In § 2.3 we propose an improved test statistic and study its asymptotic null distribution.

2.3. Improved test statistic

Under the normality assumption, given $\mathcal{X}_j^{(k)}, \hat{J}_j^{(k)}$ has the conditional normal distribution

$N\left\{t_j^{(k)}, \left(\mathcal{X}_j^{(k)T} \mathcal{X}_j^{(k)}\right)^{-1}\right\}$ for every $j > k$. This implies that, conditional on $\mathcal{X}_j^{(k)}$, the

distribution of $\hat{t}_{j,j-k}^{(k)}$ is $N(t_{j,j-k}, d_j)$ where

$$\Delta_j = \left\{ \mathcal{X}_{j-k}^T \mathcal{X}_{j-k} - \mathcal{X}_{j-k}^T \mathcal{X}_j^{(k-1)} \left(\mathcal{X}_j^{(k-1)T} \mathcal{X}_j^{(k-1)} \right)^{-1} \mathcal{X}_j^{(k-1)T} \mathcal{X}_{j-k} \right\}^{-1}. \text{ Hence}$$

$\left(\Delta_j \hat{d}_j^{(k)}\right)^{-1/2} \left(\hat{t}_{j,j-k}^{(k)} - t_{j,j-k}\right)$ follows the t -distribution with $n-k$ degrees of freedom, t_{n-k} . This result was previously noted by Wu & Pourahmadi (2003). Consequently,

$\left(\hat{t}_{j,j-k}^{(k)} - t_{j,j-k}\right)^2 / \left(\Delta_j \hat{d}_j^{(k)}\right)$ follows the F -distribution with 1 and $n-k$ degrees of freedom, $F_{1,n-k}$.

Write $\hat{t}_{j,j-k}^{(k)2} / \left(\Delta_j \hat{d}_j^{(k)}\right)$ as l_j , and modify the test statistic L to $L_c = \sum_{j=k+1}^p l_j$. For the asymptotic study of null distribution, we to further modify L_c by standardizing each l_j . Each l_j has the $F_{1,n-k}$ distribution, so the mean and variance of l_j are $E(l_j) = (n-k)/(n-k-2)$ and $\text{var}(l_j) = 2(n-k)^2(n-k-1)/\{(n-k-2)^2(n-k-4)\}$. After standardization, we get the test statistic

$$L_f = (p-k)^{-1/2} \sum_{j=k+1}^p \frac{l_j - E(l_j)}{\text{var}(l_j)^{1/2}}. \quad (3)$$

The following theorem studies the asymptotic null distribution of L_f .

Theorem 1—Under the null hypothesis $H_0 : K \leq (k-1)$, as $n, p \rightarrow \infty$, the asymptotic distribution of L_f is standard normal.

From Theorem 1 we know that for a significance level α , H_0 can be rejected if $|L_f| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of $N(0,1)$. Moreover, the asymptotic normality of L_f also holds as $p \rightarrow \infty$ with an arbitrary fixed sample size $n > k+4$.

Next, we study the power of the test using L_f . For all $k < j \leq p$, given $X_{i,j-k+1}, \dots, X_{i,j-1}$, the partial correlation coefficient between $X_{i,j}$ and $X_{i,j-k}$ is denoted by $\rho_j(k)$. One can see that $\rho_j(k) = 0$ for all $k < j \leq p$ when $H_0 : K \leq (k-1)$ is true. If $\rho_j(k)$ for some $k < j \leq p$ is large enough, many hypothesis testing methods with high power may exist; so here we only study

the power of the statistic L_f when the $\rho_j(k)$ ($k < j - p$) are all small, even tending to zero as both n and p tend to ∞ . To this end, define $\rho_{\min}(k) = \min_{k < j - p} |\rho_j(k)|$ and $\rho_{\max}(k) = \max_{k < j - p} |\rho_j(k)|$; then we have the following theorem.

Theorem 2—If $H_0 : K = (k - 1)$ is false with $K = k$, and if $n\rho_{\max}^2(k) / \{1 - \rho_{\max}^2(k)\} \rightarrow 0$ and $n\rho_{\min}^2(k) / \{1 - \rho_{\min}^2(k)\} \rightarrow \infty$ as $n, p \rightarrow \infty$, then $\text{pr}(|L_f| > z_{1-\alpha/2}) \rightarrow 1$.

Theorem 2 shows that the power of the test using our proposed statistic tends to 1 as both p and n go to ∞ . It allows $\rho_j(k)$ ($k < j - p$) to tend to zero, and the rate is controlled by the conditions $n\rho_{\min}^2(k) / \{1 - \rho_{\max}^2(k)\} \rightarrow 0$ and $n\rho_{\min}^2(k) / \{1 - \rho_{\min}^2(k)\} \rightarrow \infty$. In particular, conditions allow the $\rho_j(k)$ ($k < j - p$) to converge zero at rates faster than $n^{-1/2}$ but slower than $n^{-1/2} p^{-1/4}$. The null hypothesis $H_0 : K = (k - 1)$ is equivalent to $t_{j,j-k} = 0$ for all

$k < j - p$. For every $k < j - p$, $\left(\Delta_j \hat{d}_j^{(k)}\right)^{-1/2} \left(\hat{t}_{j,j-k}^{(k)} - t_{j,j-k}\right) \sim t_{n-k}$, which can also be used to test whether $t_{j,j-k}$ is zero (Wu & Pourahmadi, 2003). However, conditions in Theorem 2 allow the partial correlation coefficients to tend to zero, which implies that all $t_{j,j-k}$ for $k < j - p$ tend to zero. This suggests that testing whether $t_{j,j-k}$ is zero using the statistic $\left(\Delta_j \hat{d}_j^{(k)}\right)^{-1/2} \hat{t}_{j,j-k}^{(k)}$ may not have high power.

2.4. Band size detection procedure

In general, the true band size K_0 should not be too large, so it is reasonable to assume that $K \leq M$, where M is a constant smaller than $n - 4$. Here we perform a series of hypothesis tests to identify K_0 . The following sequence of hypothesis tests is considered:

$$H_{0k}: K \leq k - 1 \quad \text{versus} \quad H_{1k}: K \geq k, \quad 1 \leq k \leq M.$$

The null hypothesis H_{0k} is true only when k is larger than the actual band size K_0 . This means that $K_0 = \max\{k : H_{0k} \text{ is false}\}$. Hence we estimate K_0 as the largest k for which H_{0k} is rejected. In order to identify the actual band size, the following algorithm can be used.

Algorithm 1

Step 1: Initialization: fix the overall significance level α , and let $k = M$ where M is a prespecified upper bound on the band size.

Step 2: If $k = 0$, stop and output $\hat{K} = k$. Otherwise, following (3), compute the test statistic L_f , denoted by $L_f^{(k)}$ here. Let $p_k = 2 \left\{ 1 - \Phi \left(|L_f^{(k)}| \right) \right\}$.

Step 3: If $p_k > \alpha_k$ where α_k is the significance level for H_{0k} , then do not reject H_{0k} , update $k = k - 1$, and go to Step 2. If $p_k \leq \alpha_k$, reject H_{0k} , stop the iteration, and report the final band size estimator $\hat{K} = k$.

The value of each α_k needs to be chosen to ensure that the overall significance level of the procedure is no larger than α , and to further adjust the multiple comparison error for the whole procedure. In particular, we need to control the familywise error, i.e., the probability that H_{0k} is rejected for some $k = K_0 + 1$. There is an extensive literature on multiple comparison, and many procedures can be used; see Efron (2010) for a comprehensive review. For example, the Bonferroni procedure sets $\alpha_k = \alpha/M$. In our numerical studies in § 3, we use the Bonferroni procedure for Algorithm 1, which controls the familywise error so that it is no larger than α .

The Bonferroni procedure can be too conservative when the number of tests is relatively large. Many improved multiple comparison procedures are available; for example, Holm (1979) proposed a step-down procedure. Both the Bonferroni and the Holm procedures do not require independence of the p -values. Here the p -values are dependent, so we also consider Holm's procedure (Holm, 1979), the details of which are summarized in Algorithm 2 below.

Algorithm 2

Step 1: Initialization: fix the overall significance level α .

Step 2: For $k = 1, \dots, M$, following (3), compute the test statistic L_f , denoted by $L_f^{(k)}$ here.

Let $p_k = 2 \left\{ 1 - \Phi \left(|L_f^{(k)}| \right) \right\}$.

Step 3: Sort p_k as $p_{(1)} \dots p_{(M)}$, with the corresponding null hypotheses denoted by H_{0j1}, \dots, H_{0jM} . Reject H_{0jk} if $p_{(j)} \leq \alpha/(M - j + 1)$ for all $j = 1, \dots, k$.

Step 4: If no null hypothesis is rejected, output $\hat{K}=0$, which implies a diagonal precision matrix; otherwise, set $\hat{K} = \max \{k: H_{0k} \text{ is rejected}\}$.

Our band size detection method involves sequential testing. It will be desirable to utilize the sequential nature for multiple comparison adjustment. In a 2013 unpublished technical report (arXiv:1309.5352), M. G. G'sell et al. proposed two test procedures to handle the sequential setting. However, as shown in the Supplementary Material, these procedures perform worse than Algorithms 1 and 2 above, possibly due to the dependence of these tests.

3. Numerical Study

3.1. Simulation study

Example 1—In this example, we study the empirical size of the hypothesis test (2) using L_f . Specifically, for a given K_0 , we consider the following precision matrix: $\Omega = (\omega_{ij})$ where $\omega_{ij} = I(i=j) + \sum_{l=1}^{K_0} (2l)^{-1} I(|i-j|=l)$. Thus, the diagonal elements of Ω are all 1, and the off-diagonal elements are $(2l)^{-1}$ within a band of size K_0 and 0 elsewhere. The independent random variables X_i ($1 \leq i \leq n$) are generated from the normal distribution $N(0, \Omega^{-1})$. We set the significance level to be $\alpha = 0.05$ and use the proposed test statistic L_f to test $H_0: K \leq K_0$ versus $H_1: K > K_0 + 1$. In this case, the null hypothesis H_0 is true. For different settings of

the parameters n , p and K_0 , we repeat the test procedure 2500 times. The empirical test sizes summarized in Table 1 show that L_f can control the Type I error well.

Example 2—This example is used to demonstrate the test power of the proposed test statistic L_f . We set $K_0 = 10$ and take the same Ω as in Example 1. We consider the hypothesis test $H_0 : K = 9$ versus $H_1 : K = 10$. Since $K_0 = 10$, the alternative hypothesis is true in this case.

For different settings of the parameters n and p , we repeat the test 2500 times. The percentages of rejections of H_0 are summarized in Fig. 1(a). From the results, we see that the empirical power increases to 1 as p increases. The empirical power for $n = 100$ increases faster than that for $n = 50$. Overall, the results indicate that L_f can control the Type II error very well.

Example 3—This example looks at a different precision matrix setting. For a given K_0 , we consider the precision matrix $\Omega = (\omega_{ij})$ with $\omega_{ij} = I(i=j) + \sum_{l=1}^{K_0} 3^{-l/2} I(|i-j|=l)$. Then, the diagonal elements of Ω are 1, and the off-diagonal elements are $3^{-l/2}$ within a band of size K_0 and 0 elsewhere. The independent random variables X_i ($1 \leq i \leq n$) are generated from the normal distribution $N(0, \Omega^{-1})$. We set the upper bound of K to be $M = 8$.

For different settings of the parameters n , p and K_0 , we repeat the band detection procedure 2500 times with overall significance level $\alpha = 0.01$, and summarize the percentages of correct identification of K_0 in Table 2. The two algorithms have similar performance and can correctly identify K_0 with very high probability in most cases.

3.2. Illustration using prostate cancer data and telephone call centre data

We illustrate the application of our proposed method with a protein mass spectroscopy prostate cancer dataset previously analysed by Adam et al. (2002) and Levina et al. (2008). After performing the same preprocessing as in Levina et al. (2008), $x_i = (x_{i1}, \dots, x_{i,218})^T$ consists of an intensity profile of length 218 for the i th blood serum, with the class label y_i . There are a total of 157 healthy patients and 167 cancer patients. We apply linear discriminant analysis to classify the two groups. The dataset includes prespecified training and test sets with sample sizes $n = 216$ and $N = 108$, respectively. Since the variables have a natural order in time, we assume that the corresponding precision matrix is banded. We first use our method to detect K_0 and then use the Cholesky-based method to obtain the precision matrix estimator $\hat{\Omega}$, which is inserted into the linear discriminant rule to yield the final classifier.

We begin by calculating estimators for the precision matrix with all possible candidate values of the band size, $K = 0, 1, 2, \dots, 218$, and compute the corresponding test errors, which are plotted in Fig. 1(b). We see that when K is equal to 96 or 102, the corresponding classifier has the smallest test error, 0.1204. When the band size is larger than 150, the test error increases. Hence it is reasonable to set the upper bound of K to be $M = 150$.

Furthermore, with $\alpha = 0.01$, the final band size estimator is given as $\hat{K} = 98$ by both Algorithm 1 and Algorithm 2, with a test error of 0.1481. Although the final classifier does

not attain the minimum test error, our band size estimator is around 96 and 102. For comparison, five-fold crossvalidation using Frobenius loss is also evaluated 500 times. We count the frequencies of various candidate values of the band size selected by crossvalidation, and plot the results in Fig. 1(c). The smallest band size selected is 4, and the largest is 150. The test errors for all selected band sizes are computed, with the average test error being 0.1677. Moreover, the value selected most often by crossvalidation is 17, with a test error of 0.1574. In contrast, our proposed method is very stable and the corresponding final classifier performs well.

Our second illustration uses a call centre dataset, which consists of telephone records from a call centre of a U.S. financial institution (Huang et al., 2006; Bickel & Levina, 2008b). Phone calls were recorded from 7 a.m. until midnight on a total of 239 days in 2002. On each day, the 17-hour period was divided into 102 10-minute intervals. The number of calls arriving in each period, denoted by N_{ij} , was recorded for every day i and every 10-minute interval j . A transformation $x_{ij} = (N + 1/4)^{1/2}$ is applied to make the data closer to normal. The goal is to predict each arrival count in the second half of the day using the previous counts from that day. Specifically, we denote the mean and covariance matrix of x_i by $\mu = (\mu_1, \dots, \mu_{102})^T$ and Σ , respectively. Define $\mu^j = (\mu_1, \dots, \mu_{j-1})^T$, let $\Sigma_{(j,1:j-1)}$ be the submatrix formed by the j th row and the first $j-1$ columns of Σ , and let $\Sigma_{(1:j-1,1:j-1)}$ be the submatrix formed by the first $j-1$ rows and first $j-1$ columns of Σ . For the observation of the i th day, $x_i = (x_{i1}, \dots, x_{i,102})^T$, the best linear predictor of x_{ij} ($j = 52, \dots, 102$) from

$x_i^j = (x_{i1}, \dots, x_{i,j-1})^T$ is $\hat{x}_{ij} = \mu_j + \Sigma_{(j,1:j-1)} \left(\Sigma_{(1:j-1,1:j-1)} \right)^{-1} (x_i^j - \mu^j)$. The data were divided into a training set, consisting of data from the first 205 days, and a test set, comprising data from the last 34 days. We first use our method to detect the band size K_0 of the precision matrix, and then use the Cholesky-based method to obtain the estimator for the covariance matrix Σ of the above best linear predictor. The performance is measured by the average error $\sum_{j=52}^{102} E_j / 51$, where $E_j = \sum_{i=206}^{239} |\hat{x}_{ij} - x_{ij}| / 34$. With $\alpha = 0.01$ and the upper bound $M = 100$ on band size, the final band size estimator is given as $\hat{K} = 8$ and $\hat{K} = 10$ by Algorithms 1 and 2, with average errors 0.5584 and 0.5723, respectively. Bickel & Levina (2008b) selected the band size as 19, with an average error of 0.5809.

Supplementary Material

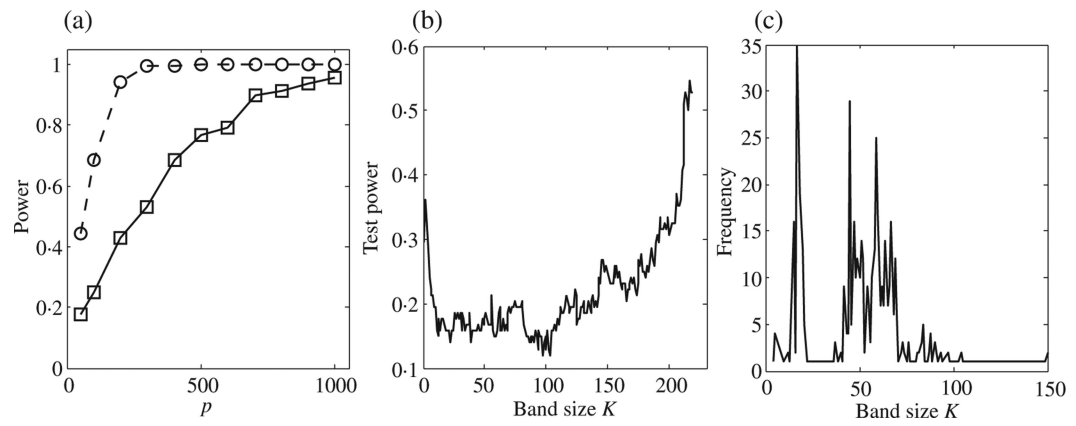
Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank the editor, associate editor and two reviewers for their helpful suggestions. An and Liu were partially supported by the U.S. National Institutes of Health and National Science Foundation. Guo was partially supported by the Program for Changjiang Scholars and Innovative Research Team in University, the National Natural Science Foundation of China, and the Jilin Project. This work was completed during An's visit to the Department of Statistics and Operations Research at the University of North Carolina. Liu is also a faculty member of the Carolina Center for Genome Sciences and the Department of Biostatistics at the University of North Carolina.

References

- Adam B, Qu Y, Davis J, Ward M, Clements M, Cazares L, Semmes O, Schellhammer P, Yasui Y, Feng Z, Wright G. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* 2002; 62:3609–14. [PubMed: 12097261]
- Bickel PJ, Levina E. Covariance regularization by thresholding. *Ann. Statist.* 2008a; 36:2577–604.
- Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Ann. Statist.* 2008b; 36:199–227.
- Cai TT, Zhang C, Zhou H. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* 2010; 38:2118–44.
- Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press; Cambridge: 2010.
- Friedman J, Hastie TJ, Tibshirani RJ. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics.* 2008; 9:432–41. [PubMed: 18079126]
- Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 1979; 6:65–70.
- Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika.* 2006; 93:85–98.
- Levina E, Rothman A, Zhu J. Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Statist.* 2008; 2:245–63.
- Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* 2006; 34:1436–62.
- Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika.* 1999; 86:677–90.
- Rothman A, Levina L, Zhu J. Generalized thresholding of large covariance matrices. *J. Am. Statist. Assoc.* 2009; 104:177–86.
- Rothman A, Levina L, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika.* 2010; 97:539–50.
- Wu WB, Pourahmadi M. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika.* 2003; 90:831–44.
- Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika.* 2007; 94:19–35.

**Fig. 1.**

Plots for the numerical studies: (a) empirical power of L_f for Example 2 with $n = 50$ (solid line) and $n = 100$ (dashed line); (b) test errors for various values of K for the cancer example; (c) frequencies of various candidate values of K selected by five-fold crossvalidation for the cancer example.

Table 1
Empirical test sizes (%) for Example 1 based on 2500 simulation replications; here the true α is 5%

<i>n</i>	<i>p</i>					<i>p</i>				
	30	100	200	500	1000	30	100	200	500	1000
	<i>K</i> ₀ = 2					<i>K</i> ₀ = 4				
50	4.2	5.4	4.3	3.8	5.0	4.2	5.0	4.0	5.1	4.6
200	5.0	4.8	4.8	4.4	5.6	5.0	4.2	5.2	5.0	4.8
400	3.9	4.6	4.4	5.0	5.0	4.5	5.0	5.3	4.5	4.7

Table 2

Percentages (%) of correct identifications of K_0 in Example 3

		p					p				
		$K_0 = 2$					$K_0 = 6$				
	n	30	100	200	500	1000	30	100	200	500	1000
Algorithm 1	50	96.9	99.0	98.9	99.3	99.1	1.8	2.8	6.2	15.9	42.3
	200	97.8	98.7	99.1	99.5	99.3	20.0	70.1	96.8	99.8	99.6
	400	97.8	98.7	98.9	99.4	99.2	65.6	99.4	99.8	99.6	99.7
Algorithm 2	50	96.9	98.4	98.6	99.1	99.1	2.4	3.4	7.3	20.6	49.2
	200	97.2	98.2	98.4	99.2	98.8	23.1	76.2	97.9	98.9	98.9
	400	97.8	97.9	99.2	98.7	98.9	69.8	98.8	98.8	98.8	98.9